Using Embeddings to Understand the Variance and Evolution of Data Science Skill Sets

Maryam Jahanshahi Ph.D.
Research Scientist
TapRecruit.co

Jobs are hard to categorize

Research Analyst — Entry-Level

DRG - New York City

We are seeking a Research Analyst to join our team which serves our pharmaceutical clients with actionable data. As a Research Analyst, you will provide data support for client-facing platforms, presentations, and client requests…

Research Analyst — Senior

MMP - New York City

We have an exciting opportunity for an individual to join MMP's Cyber Risk Group. The successful candidate will have the ability to shape our investors service strategy, analytic, research and outreach framework for cyber risk and its relationship to credit and the financial markets…

Research at TapRecruit

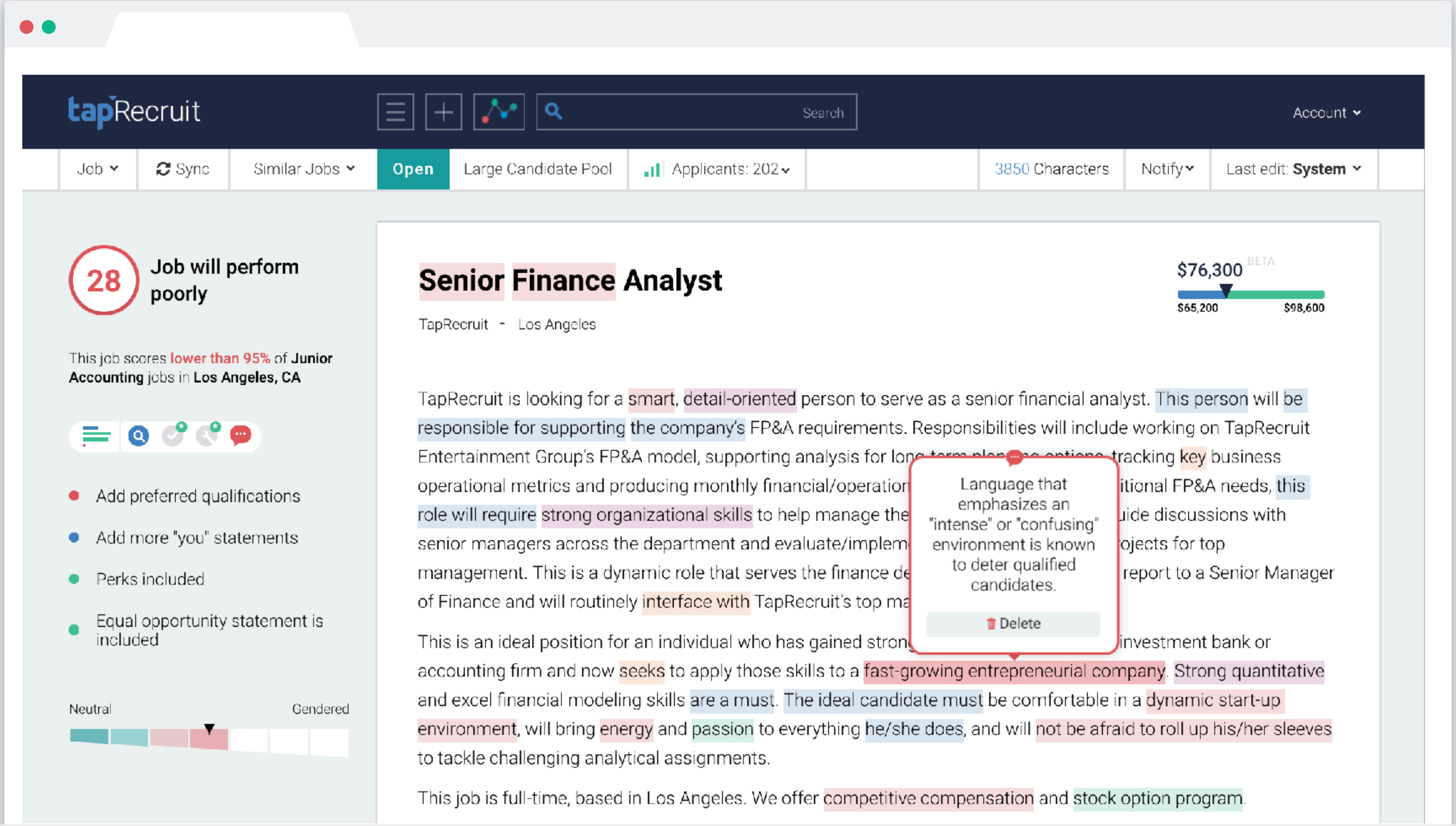
Helping companies make fairer and more efficient recruiting decisions

NLP and Data Science:

- What are distinguishing characteristics of successful career documents?
- What skills are increasingly important for different industries?

Decision Science:

- How do candidates make decisions about which jobs to apply to?
- How do hiring teams make decisions about candidate qualifications?


tapRecruit
Search
Account
Job
Sync
Similar Jobs
Open
Large Candidate Pool
Applicants: 202
3850 Characters
Notify
Last edit: System
28
Job will perform poorly
This job scores lower than 95% of Junior Accounting jobs in Los Angeles, CA
Add preferred qualifications
Add more "you" statements
Perks included
Equal opportunity statement is included
Neutral
Gendered
Senior Finance Analyst
TapRecruit - Los Angeles
$76,300
BETA
$65,200
$98,600
TapRecruit is looking for a smart, detail-oriented person to serve as a senior financial analyst. This person will be responsible for supporting the company's FP&A requirements. Responsibilities will include working on TapRecruit Entertainment Group's FP&A model, supporting analysis for lon
tracking key business
operational metrics and producing monthly financial/operation
itional FP&A needs, this role will require strong organizational skills to help manage the
ide discussions with senior managers across the department and evaluate/implem
ojects for top management. This is a dynamic role that serves the finance de
report to a Senior Manager of Finance and will routinely interface with TapRecruit's top ma
Language that emphasizes an "intense" or "confusing" environment is known to deter qualified candidates.
Delete
This is an ideal position for an individual who has gained stron
investment bank or accounting firm and now seeks to apply those skills to a fast-growing entrepreneurial company. Strong quantitative and excel financial modeling skills are a must. The ideal candidate must be comfortable in a dynamic start-up environment, will bring energy and passion to everything he/she does, and will not be afraid to roll up his/her sleeves to tackle challenging analytical assignments.
This job is full-time, based in Los Angeles. We offer competitive compensation and stock option program.

How have data science skills changed over time?

Strategies to identify changes in texts

Traditional approaches do not capture syntactic and semantic shifts

MBA	**SQL**
PhD	**Tableau**
Python	**PowerBI**

1880	1920	1960	2000
force	**atom**	**radiat**	**state**
energy	**theory**	**energy**	**energy**
motion	**electron**	**electron**	**electron**
differ	**energy**	**measure**	**magnet**
light	**measure**	**ray**	**field**

Matter

Quantum

Electron

Manual Feature Extraction

Require selection of key attributes, therefore difficult to discover new attributes

Dynamic Topic Models

Require experimentation with topic number

Adapted from Blei and Lafferty, ICML 2006.

Embeddings use context to extract meaning

Window sizes capture semantic similarity vs semantic relatedness

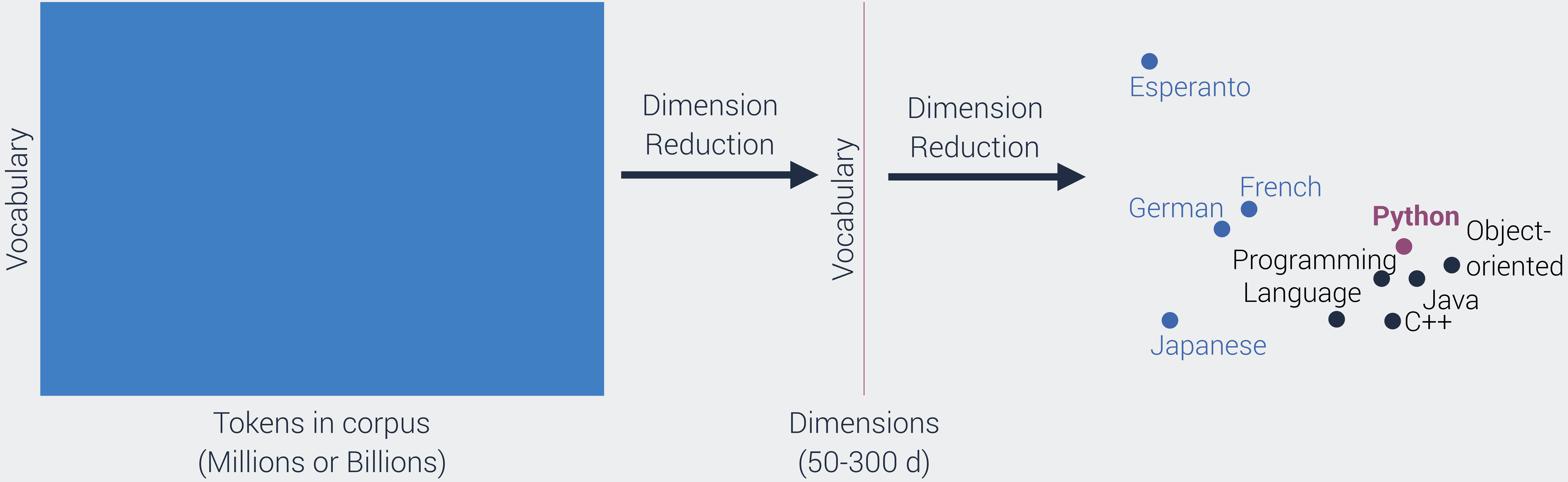
$\underbrace{\text{Statistical modeling through software (e.g. SPSS) or programming language (e.g.}}_{\textbf{Context}}$ $\underbrace{\textbf{Python}}_{\textbf{Word}}$)

$\underbrace{\text{Experience in}}_{\textbf{Context}}$ $\underbrace{\textbf{Python}}_{\textbf{Word}}$, $\underbrace{\text{Java or other object-oriented programming languages}}_{\textbf{Context}}$

$\underbrace{\text{Proficiency programming in}}_{\textbf{Context}}$ $\underbrace{\textbf{Python}}_{\textbf{Word}}$, $\underbrace{\text{Java or C++.}}_{\textbf{Context}}$

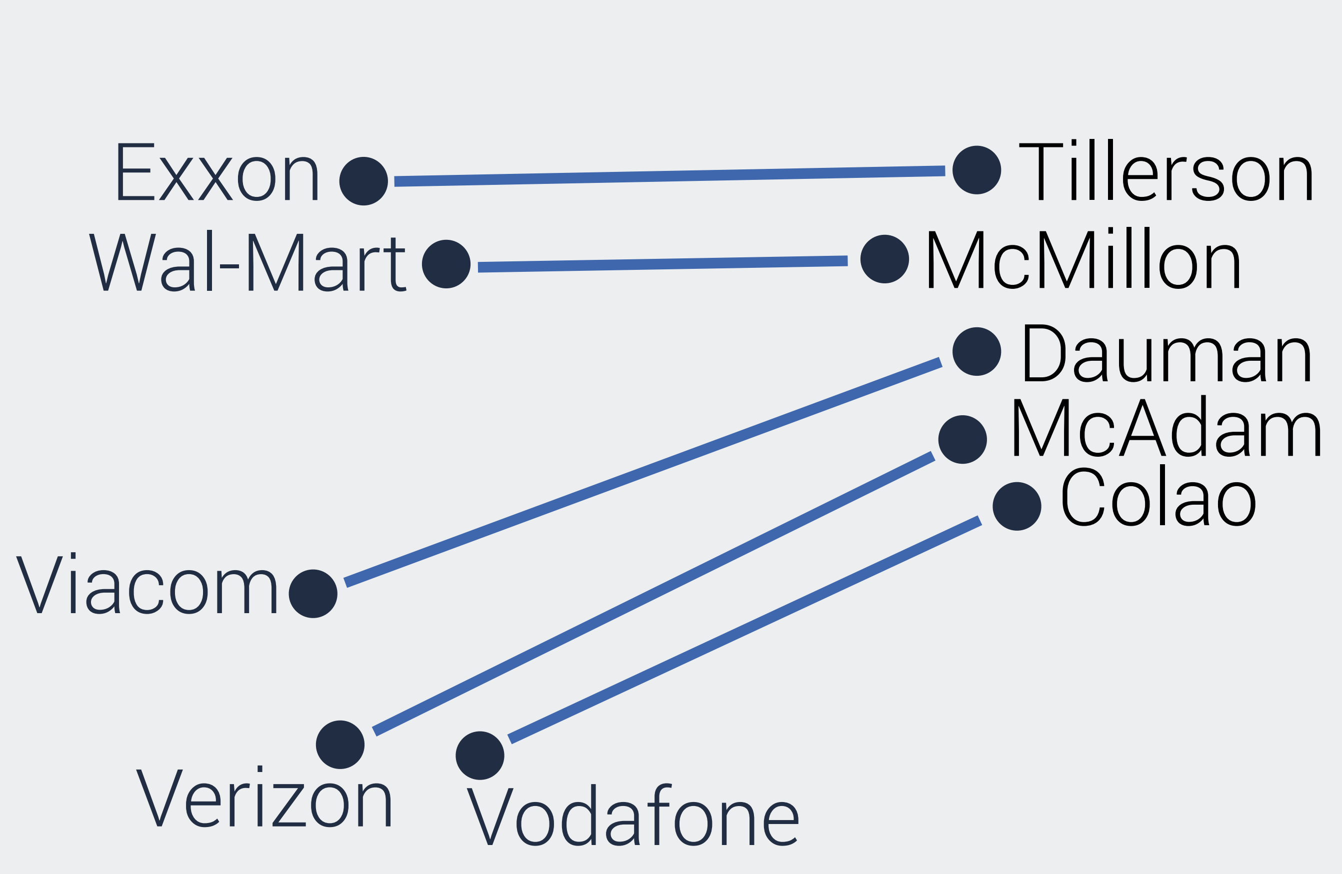
A simplified representation of word vectors

Dimension reduction is key to all types of embeddings models

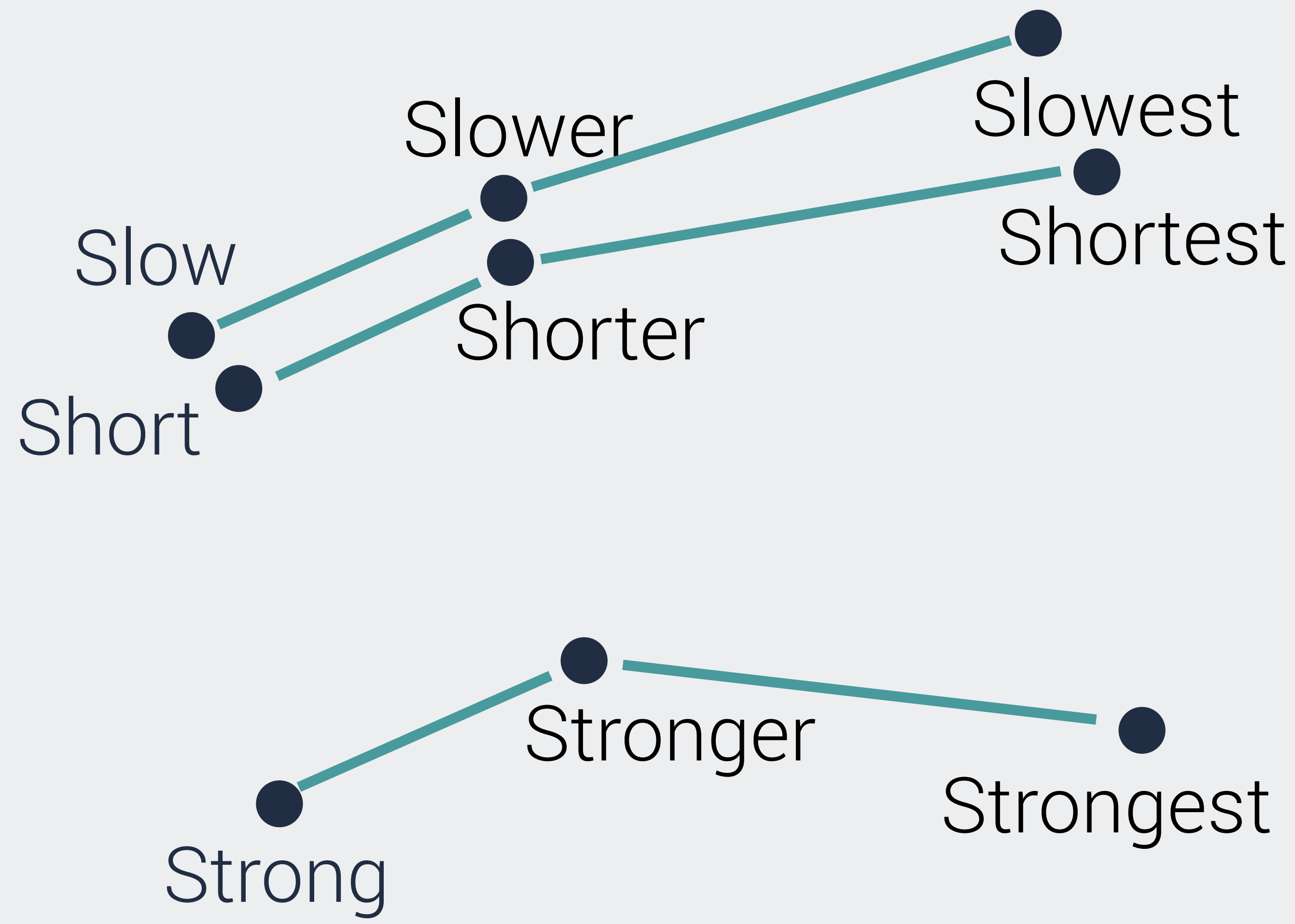


Embeddings capture entity relationships

Dimensionality enables comparison between word pairs along many axes



Hierarchies



Comparatives and Superlatives

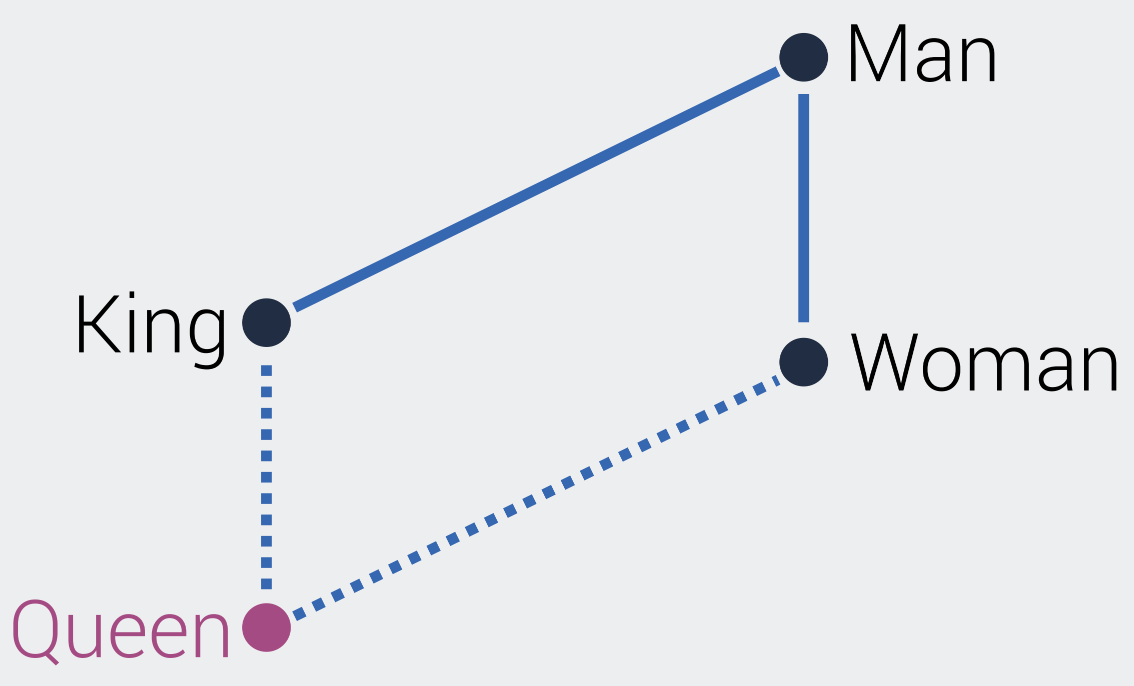
Man :: King as Woman :: ?

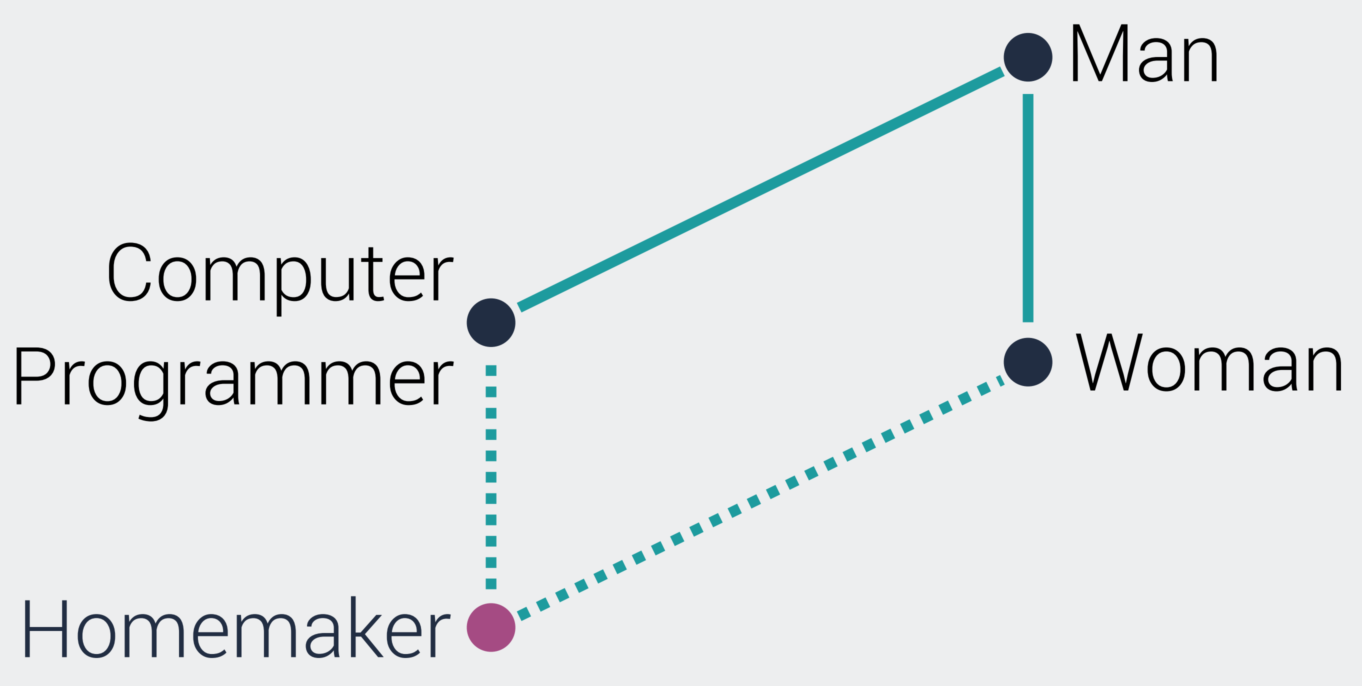
Adapted from Stanford NLP GLoVE Project

tapRecruit.co

Embeddings reflect cultural bias in corpora

High dimensionality enables some bias reduction





Man :: King as Woman :: ?

Man :: Programmer as Woman :: ?

Adapted from Bolukbasi et al., arXiv: 1607:06520.

Pretrained embeddings facilitate fast prototyping

Embeddings training should match corpus that is being tested on

- Corpus Generation
- Corpus Processing
- Language Model Generation
- Language Model Tuning
- Final Application

Corpus	Twitter	Common Crawl	GoogleNews	Wikipedia
Tokens	27 B	42-840 B	100 B	6 B
Vocabulary Size	1.2 M	1.9-2.2 M	3 M	400 k
Algorithm	GLoVE	GLoVE	word2vec	GLoVE
Vector Length	25 - 200 d	300 d	300 d	50 - 300 d

Problems with pretrained embedding models

Casing	Abbreviations vs Words e.g. IT vs it
Out of Vocabulary Words	Domain Specific Words & Acronyms
Polysemy	Words with multiple meanings e.g. drive (a car) vs drive (results) e.g. Chef (the job) vs Chef (the language)
Multi-word Expressions	Phrases that have new meanings e.g. Front-end vs front + end

Custom language models tools

Modularized for different data and modeling requirements

spaCy OpenNLP

CoreNLP SyntaxNet

gensim PYTORCH

TensorFlow Amazon SageMaker

Corpus Processing

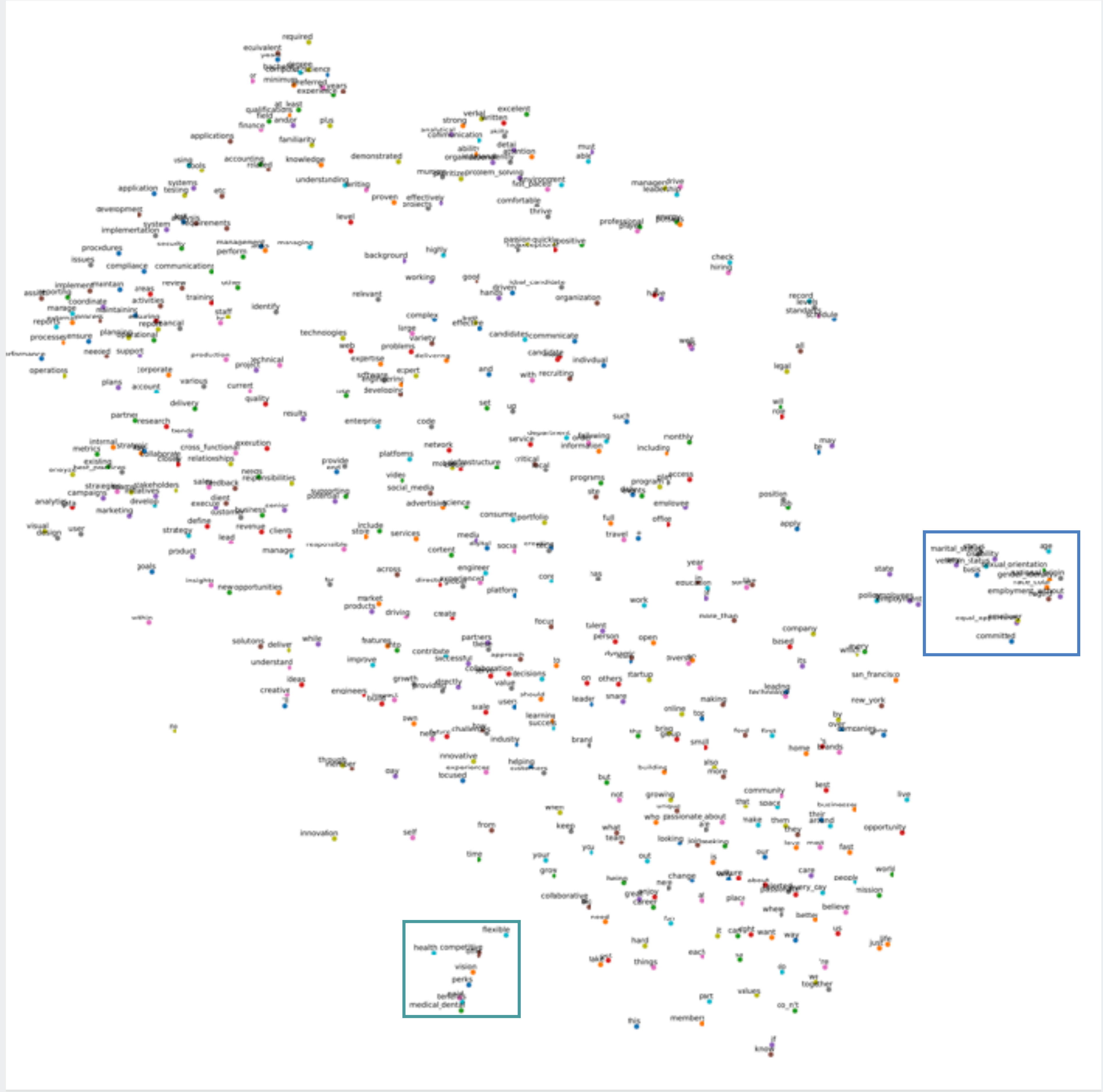
Tokenization, POS tagging, Sentence Segmentation, Dependency Parsing

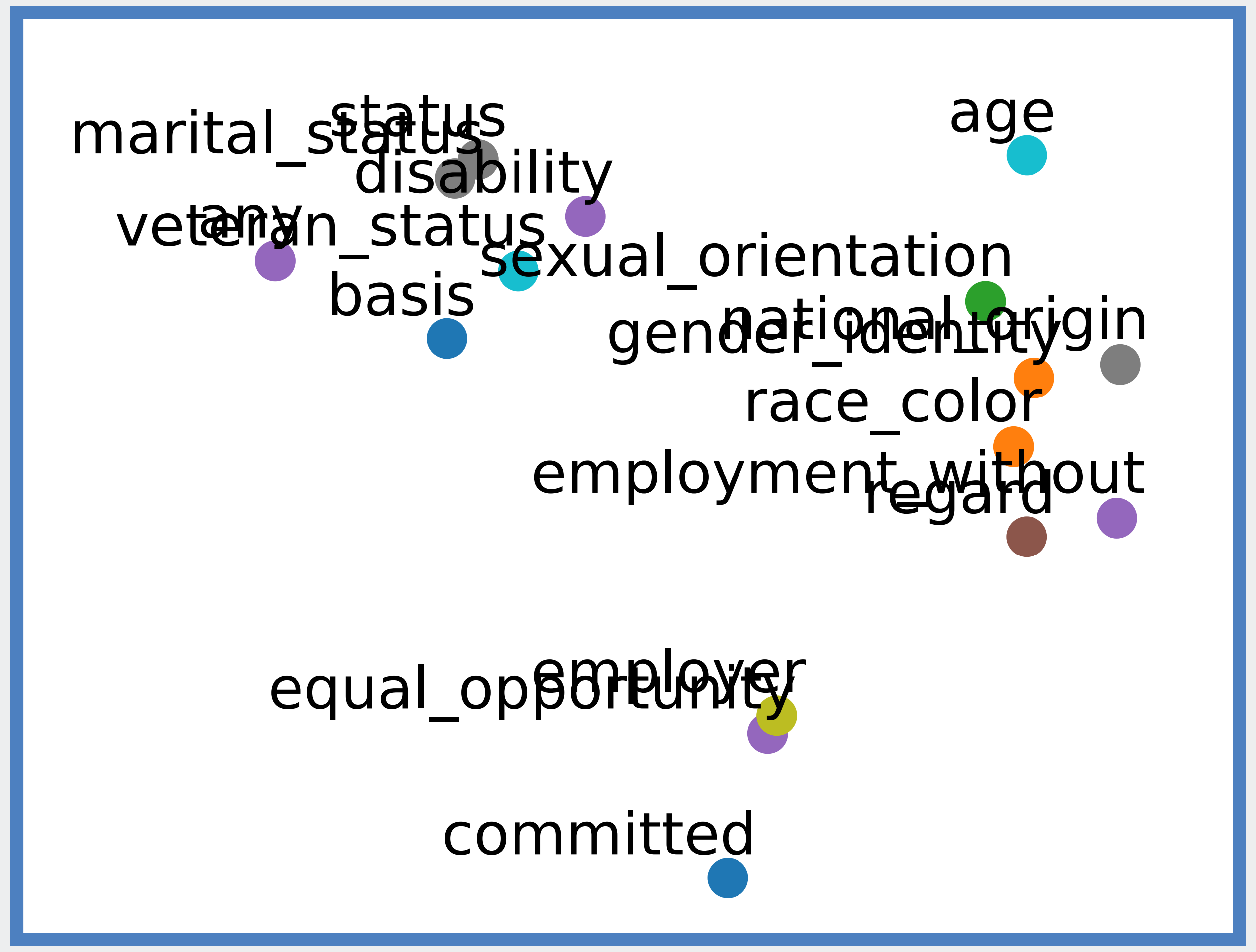
Language Modeling

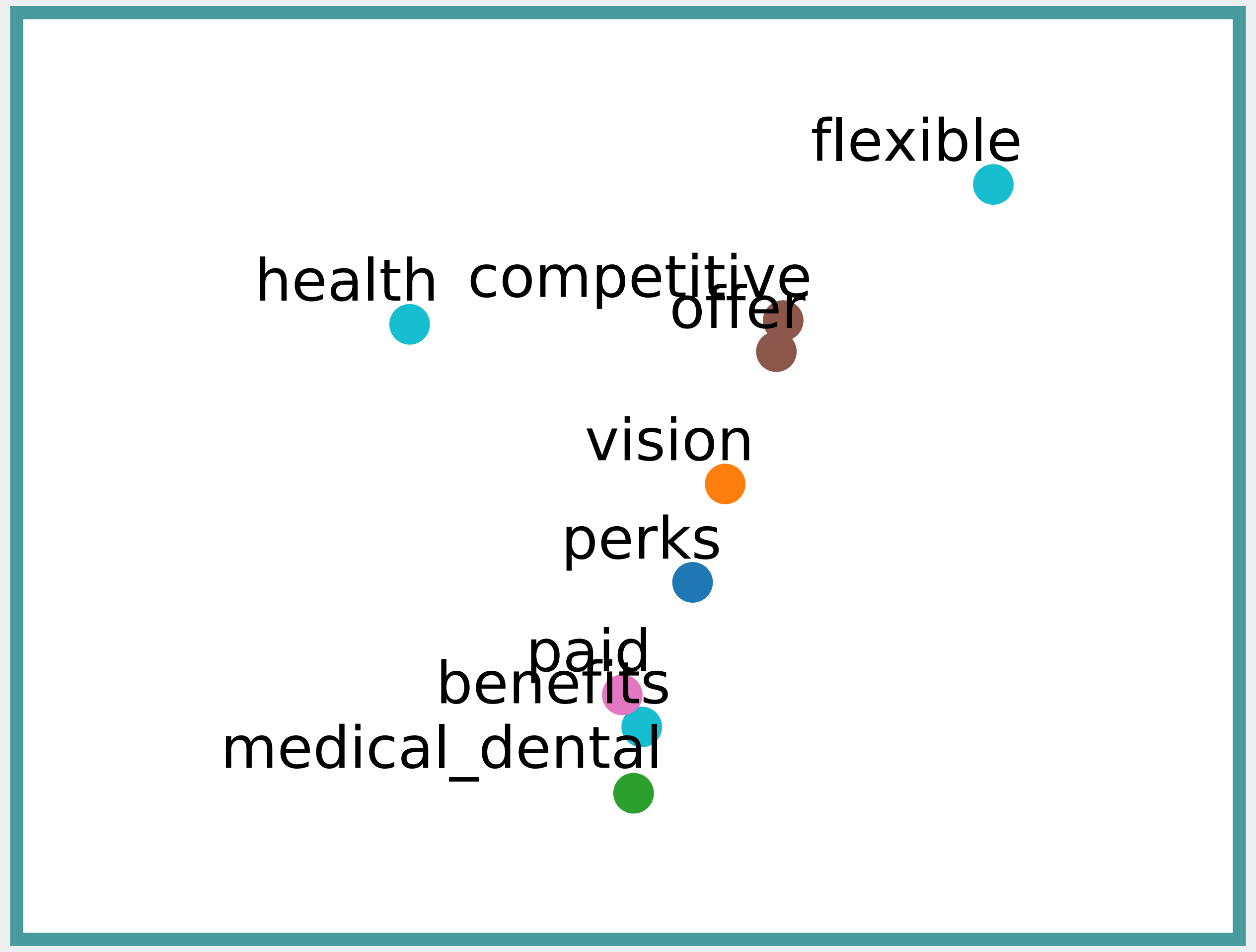
Different word embedding models (GLoVE, word2vec, fastText)

Career language embedding model

Identified equal opportunity and perks language



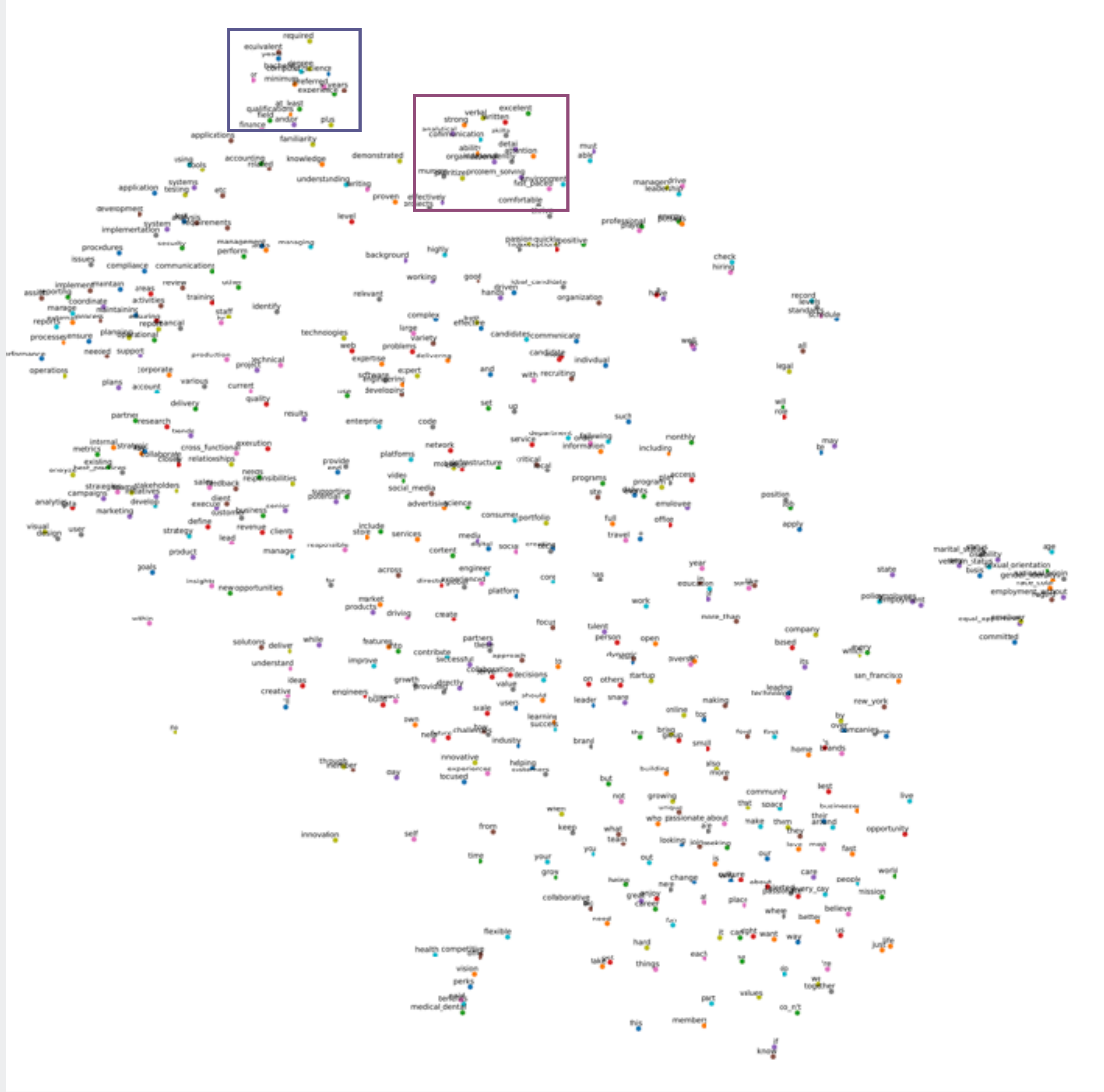


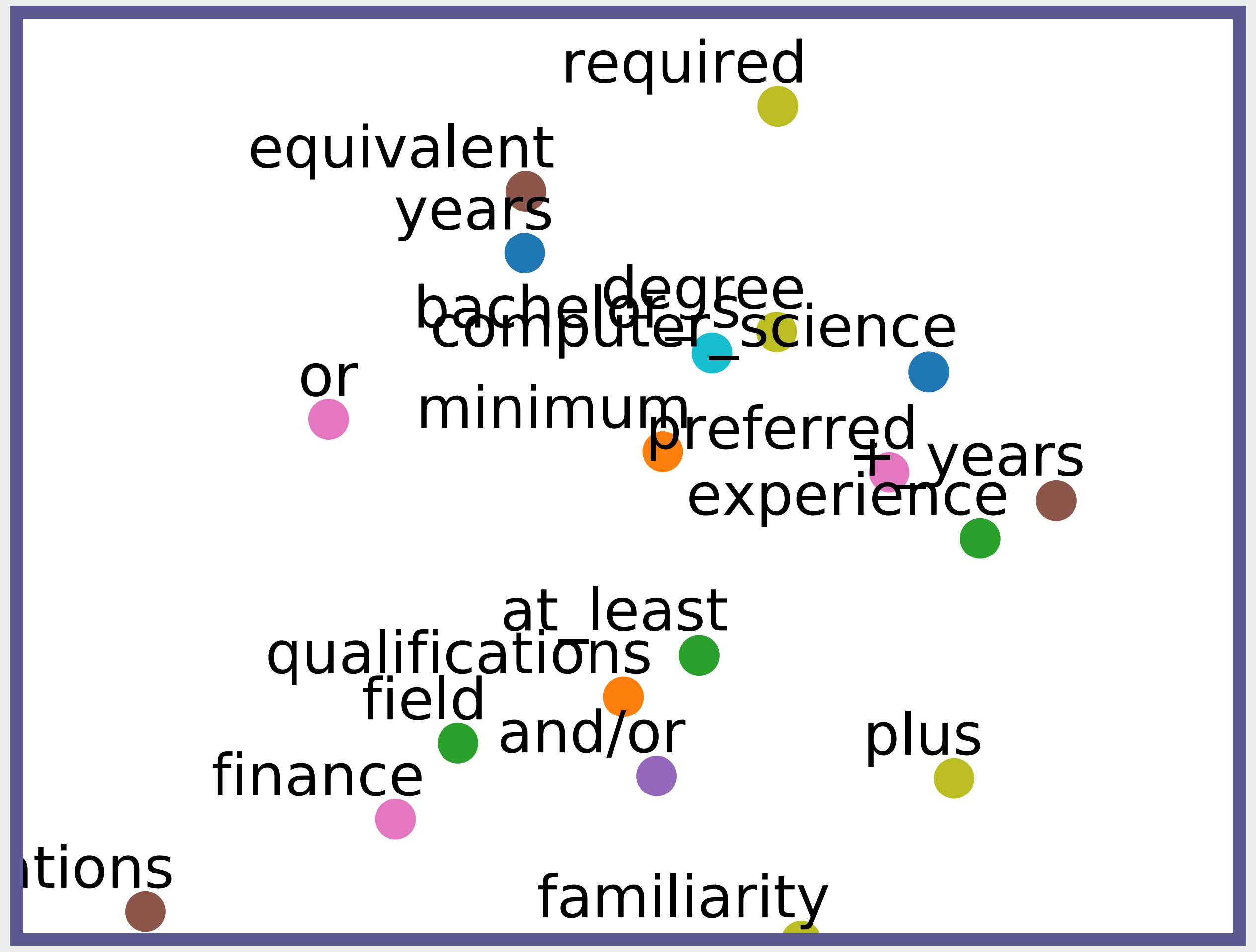


tapRecruit.co

Career language embedding model

Identified 'soft' skills and language around experience



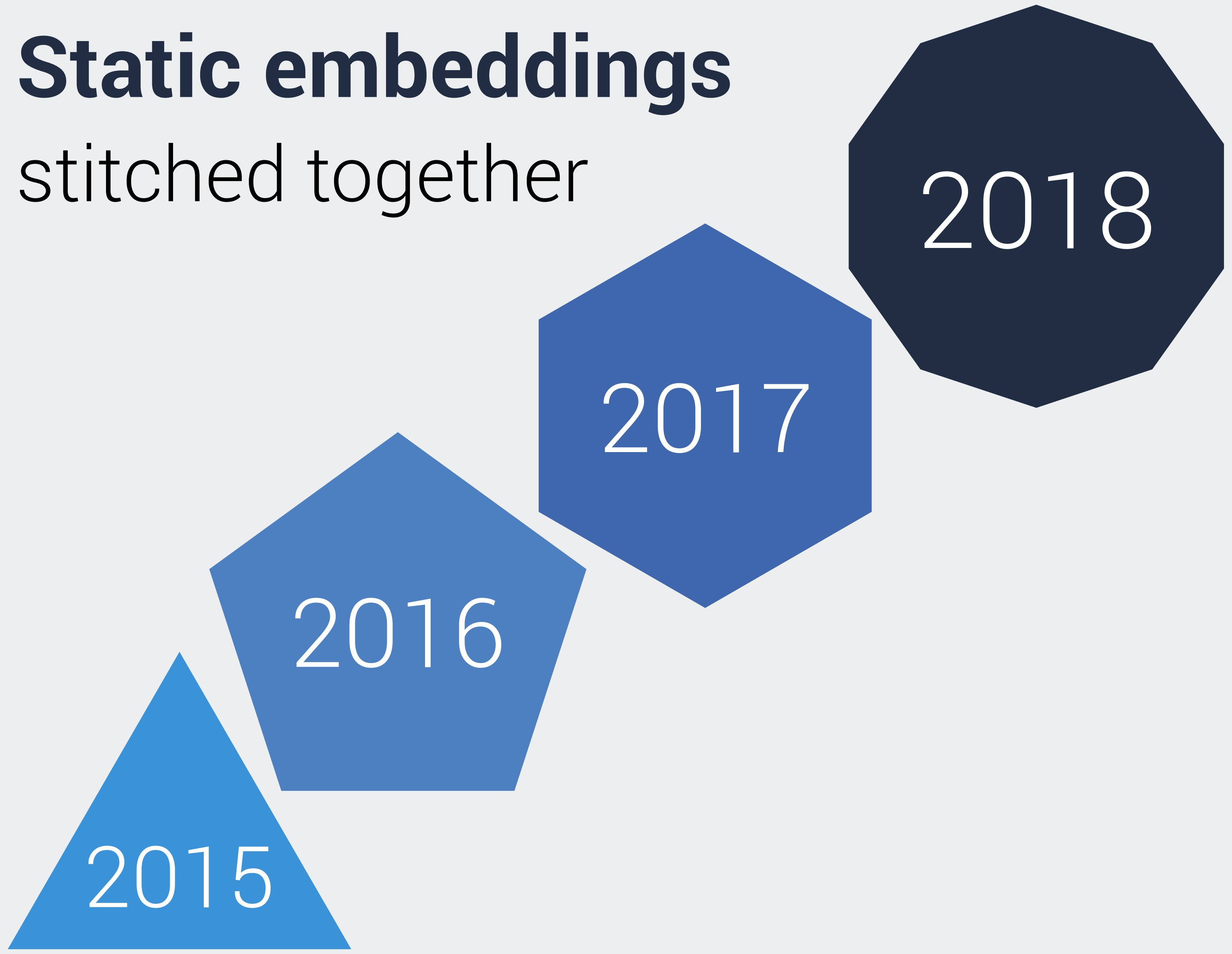


tapRecruit.co

I’ve got 300 dimensions…
but time ain’t one

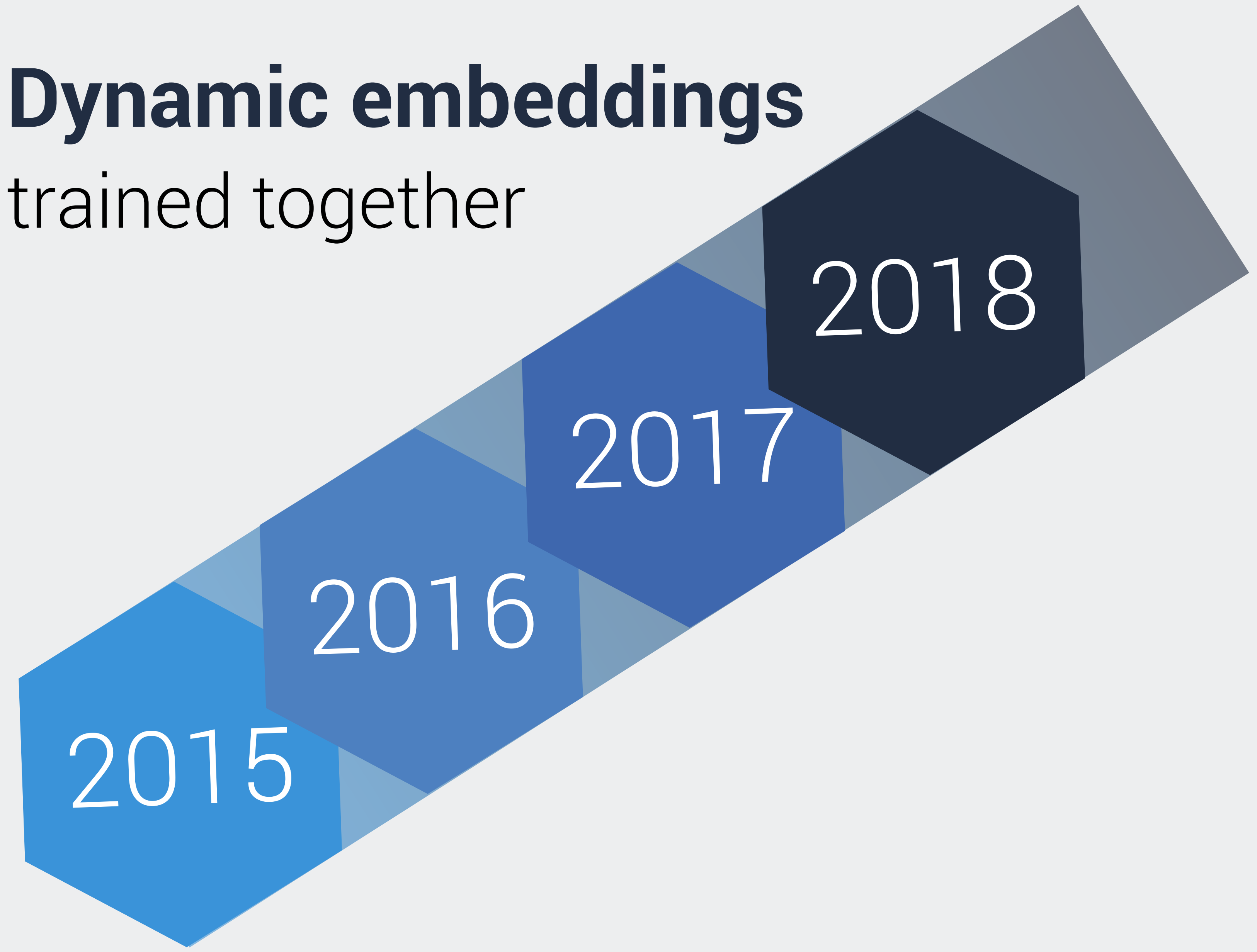
Two approaches to connect embeddings

Static embeddings
stitched together



Data hungry
Requires alignment

Dynamic embeddings
trained together



Data efficient
Does not require alignment

Kim, Chiu, Kaneki, Hedge and Petrov, arXiv: 1405:3515.
Kulkarni, Al-Rfou, Perozzi and Skiena, arXiv: 1411:3315.

Balmer and Mandt, arXiv: 1702:08359
Yao, Sun, Ding, Rao and Xiong, arXiv: 1703:00607
Rudolph and Blei, arXiv: 1703:08052

Dynamic Bernoulli embeddings

Outputs facilitate quick analysis of trends

Absolute drift

Identifies top words whose usage changes over time course

words with largest drift (Senate)			
IRAQ	3.09	coin	2.39
tax cuts	2.84	social security	2.38
health care	2.62	FINE	2.38
energy	2.55	signal	2.38
medicare	2.55	program	2.36
DISCIPLINE	2.44	moves	2.35
text	2.41	credit	2.34
VALUES	2.40	UNEMPLOYMENT	2.34

Embedding neighborhoods

Extract semantic changes by nearest neighbors of drifting words

UNEMPLOYMENT		
1858	**1940**	**2000**
unemployment	unemployment	unemployment
unemployed	unemployed	jobless
depression	depression	rate
acute	alleviating	depression
deplorable	destitution	forecasts
alleviating	acute	crate
destitution	reemployment	upward
urban	deplorable	lag
employment	employment	economists
distressing	distress	predict

Repository Link: http://bit.ly/dyn_bern_emb

Experiments with dynamic embeddings

	Small Corpus	Large Corpus
Job Types	All	All
Time Slices	3 (2016-2018)	3 (2016-2018)
Number of Documents	**50 k**	**500 k**
Vocabulary Size	10 k	10 k
Data Preprocessing	Basic	Basic
Embedding Dimensions	100 d	100 d

Dynamic embeddings

Small corpus identified gains and losses

Demand for PhDs and MBAs is Falling

PhDs in All Jobs
-23%

PhDs in DS Jobs
-30%

MBAs in All Jobs
-35%

Data Science skills showing significant shifts

Tableau
+20%

Spark
Steady

Python
Steady

PowerBI
+100%

Hadoop
-30%

Perl
-40%

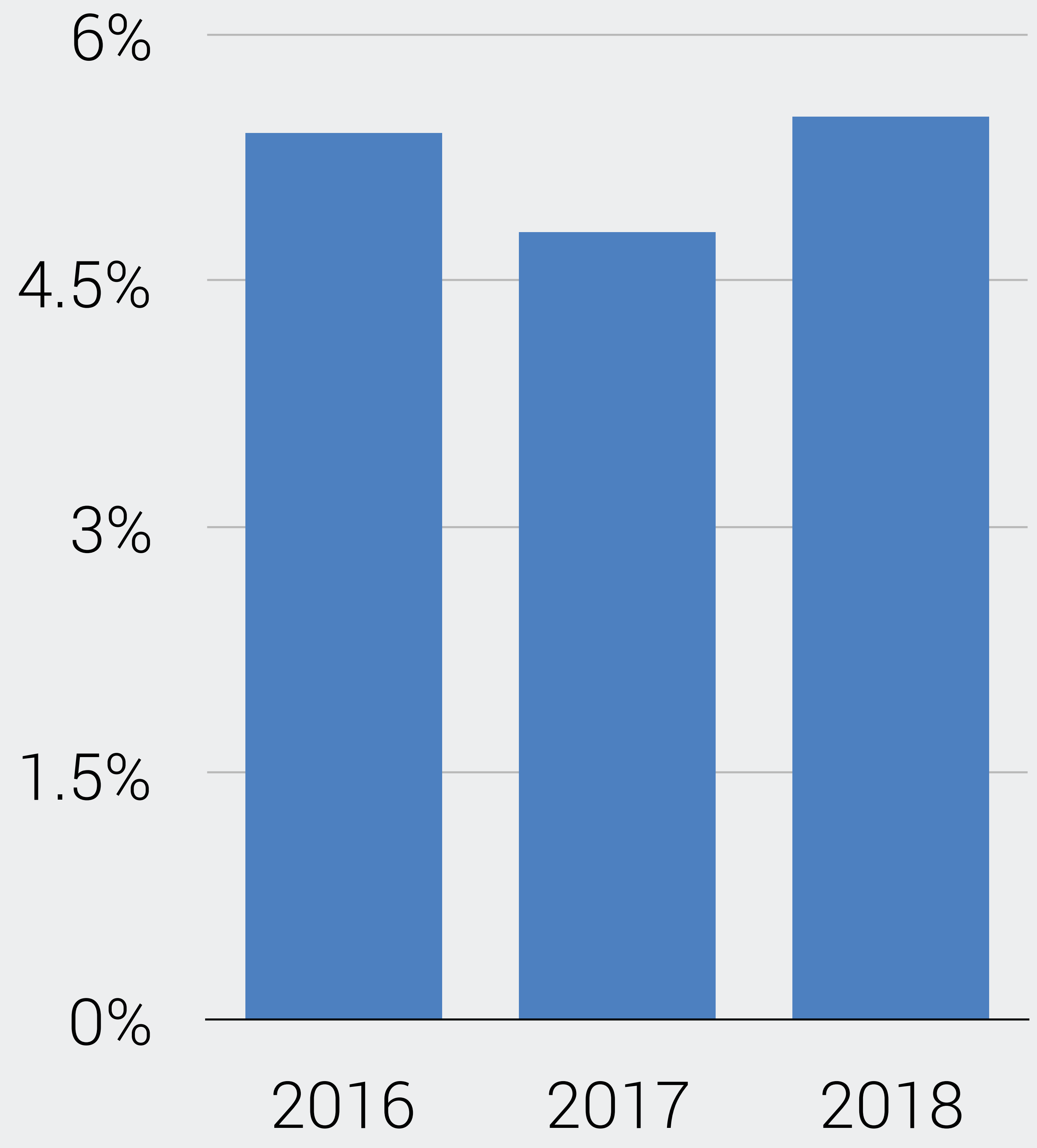
Blue boxes indicate phrases identified from top drifting words analysis.
Grey boxes indicate 'control' skills.

tapRecruit.co

Dynamic embeddings

Large corpus identified role-type dependent shifts in requirements

No change to SQL demand



SQL requirement increases in specific functions

FP&A Roles	Sales Roles	Marketing Roles
+70%	**Steady**	**Steady**
FinTech Roles	**BizDev Roles**	**HR Roles**
Steady	**+50%**	**+25%**

Blue boxes indicate phrases identified from top drifting words analysis.
Grey boxes indicate ‘control’ skills.

tapRecruit.co

How have data science skills changed over time?

- Flavors of static word embeddings: The Corpus Issue
- Considerations for developing custom embedding models
- Flavors of dynamic models: Dynamic Bernoulli embeddings

Thank you Women in Analytics!

Maryam Jahanshahi Ph.D.
Research Scientist
@mjahanshahi
maryam-j

tapRecruit.co

http://bit.ly/wia-2019